Songklanakarin J. Sci. Technol.
45 (6), 638–645, Nov. – Dec. 2023

http://www.sjst.psu.ac.th

Original Article

Enhancing the accuracy of tropospheric ozone prediction using probability distribution

Muhammad Ismail Jaffar[1], Hazrul Abdul Hamid[2*], Riduan Yunus[1], and Ahmad Fauzi Raffee[1]

[1] *Faculty of Civil Engineering and Built Environment, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, 86400 Malaysia*

[2] *Division of Mathematics, School of Distance Education, Universiti Sains Malaysia, Penang, 11800 Malaysia*

Received: 18 July 2023; Revised: 5 October 2023; Accepted: 6 November 2023

Abstract

Tropospheric ozone or ground-level ozone, mainly found near ground level, has adverse effects on human health. Distribution fitting is useful for predicting the probability, or forecasting the frequency of recurrence, of a phenomenon in a specific period of time. This study aimed to find the best fit distribution of ground-level ozone for specific industrial, rural, and suburban areas of monitoring locations in Malaysia, which were Kuala Terengganu, Jerantut, and Banting. Secondary data from 2017 to 2020 used in this study were obtained from the Department of Environment Malaysia (DoE). This study employed eight probability distributions namely Weibull, gamma, lognormal, logistic, log-logistic, Birnbaum–Saunders, Nakagami, and inverse Gaussian. The method of moments was used to estimate the parameters for each distribution and the best distribution can be used for predicting the return period of the concentration. The descriptive statistics analysis showed that ground-level ozone reached the highest peak at 1400 and 1500 hours, due to the UV radiation from sunlight, while the lowest concentration reading was at 0700 hours at all monitoring locations. By comparing the analysis of the eight distributions, Nakagami was found to be the best fit distribution to the actual monitoring data for Kuala Terengganu, Jerantut, and Banting stations from 2017 to 2020. As a result, this study suggests that the Nakagami distribution be used to predict exceedances and return periods, based on the performance indicators. Thus, it can take the place of the typical distributions employed in fitting the distribution of air pollutants, such as the lognormal distribution and the gamma distribution.

Keywords: air pollution modelling, tropospheric ozone, return period, statistical distribution

1. Introduction

Ozone has emerged as a serious pollutant in megacities and rapidly growing countries. It has the potential to significantly impact human health, vegetation, and climate (Zeng *et al.*, 2018). When nitrogen oxides (NO_x) react with oxygen molecules (O_2) in the presence of solar light, they produce free oxygen atoms (O), which combine to form ozone (O_3) (Awang & Ramli, 2017). Several factors influence the O_3 concentration, including cloud cover, sunlight, NOx, carbon monoxide (CO), and volatile organic compounds (VOC) via reactions (Latif, Huey, & Juneng, 2012). In addition, particulate matter and meteorological variables also affect the reading of O_3 (Zhou, Cheng, Zhang, Wang, & Yang, 2022).

Furthermore, O_3 concentrations are always higher in the afternoon due to the amount of UV radiation from sunlight. The two causes of ambient air pollution are man-made activities (anthropogenic) and natural sources. Natural sources include pollen dispersal, forest fires, and windblown dust (Mabahwi, Ling, & Omar, 2014). Surprisingly, about 1.2 million deaths due to air pollution were estimated in China in 2010, accounting for 35% of all such deaths worldwide

*Corresponding author
Email address: hazrul@usm.my

(Zhang *et al.*, 2018). According to Kim, Kabir, and Kabir (2015), the most hazardous air pollutant to human health is O_3, which kills 0.47 million people worldwide. Furthermore, numerous studies have indicated that the ozone pollutant has a negative impact on human health. Bolsoni, Oliveira, Pedrosa, and Souza (2018) discovered that ozone pollutant has an aggressive action on vegetation, entering the plants through stomata. Once inside the stomatal complex, it becomes soluble when it molecularly diffuses into the subestomatic cavity by spreading the intercellular gap of the mesophyll. This has occurred as a result of changes in VOC levels.

Previous studies have discovered that O_3 concentrations are generally higher during the day and lower at night and in the early morning (Song & Hao, 2015). As the ozone pollutant increases, forecasting and modeling air pollution can be used to identify and provide a tool to evaluate the ozone concentration in the future. Statistical analysis is commonly utilized to analyze existing and future air quality, particularly with regard to ozone pollutants. Maciejewska, Rezlar, Reizer, and Klejnowski in 2015 studied modeling of black carbon (BC) concentration in Warsaw, Poland by using statistical distribution and return period of the extreme concentration. Lognormal, Weibull, and gamma distributions were used, and the lognormal distribution was found to be the most appropriate to represent the middle-range values.

Previous research on O_3 concentration features and distribution fitting, on the other hand, has only looked at parent distributions such as lognormal, Weibull, and gamma. This study is more in-depth because it contains a features analysis as well as the fitting of eight different types of statistical distributions. The goal of this study was to find the best fit probability distribution and compare it to the parent distributions in air quality modelling in order to predict O_3 concentration exceedances and return periods at three monitoring locations: Jerantut, Kuala Terengganu, and Banting. To identify the best model, prediction values for each statistical distribution will be compared using five types of performance indicators: root mean square error (RMSE), normalized absolute error (NAE), index of agreement (IA), coefficient of determination (R^2), and prediction accuracy (PA).

2. Materials and Methods

2.1 Area of study and data

Three locations were selected for this study, Kuala Terengganu, Jerantut, and Banting. Kuala Terengganu monitoring station is located in the East Coast of Peninsular Malaysia, facing the South China Sea (05º18.455'N, 103º07.213'E). This monitoring station is categorized as an urban area by Department of Environment Malaysia, and surrounded by a high-density population, an airport, commercial areas, and congested traffic (Ahmad Isiyaka *et al.*, 2014). Nevertheless, this monitoring station is influenced by the inter monsoons, South-West monsoon, and North-East monsoon (Abdullah, Ismail, Yuen, Abdullah, & Elhadi, 2017). Meanwhile, the Jerantut monitoring station is located at Batu Embun, Pahang (03º55.59'N, 102º22.120'E). It is located approximately 180 km from Kuantan and 200 km from Kuala Lumpur, capital city of Malaysia. Jerantut monitoring station is situated in a rural area that has a low potential of air pollution concentration (Mohammad, Deni, & Ul-Saufie, 2018) and is surrounded by the forest as well as villages and agricultural areas (Zaki, Faizah, Yusof, & Shith, 2016). This station is considered a reference site since it is located in a rural area near a conserved forest. The Banting monitoring station is located at Kolej Mara Banting, Bukit Changgang (2.83º N, 101.62ºE). This monitoring station is categorized as a suburban area and surrounded by villages, residential areas, and palm oil estates (Suparta, Alhasa, Singh, & Latif, 2015). This study relies on secondary data supplied by the Department of Environment. The data quality has been ascertained through verification procedures in line with the department's instrumentation standards. Nevertheless, it's important to note that each dataset contains some missing values. To address these gaps, a mean top-bottom method has been employed for imputing the missing data points. The ground level ozone (O_3) hourly average concentration data from these three monitoring stations were used in this study to fit eight types of distributions, including the lognormal, gamma, and Weibull distributions, which are the three main distributions that are most frequently used in Malaysia to approximate the distribution of air pollutants. These data sets span four years, from January 2017 to December 2020. A map of the monitoring sites used in this study is shown in Figure 1.

Figure 1. Locations of the monitoring stations

2.2 The distribution

The three parent distributions that are commonly used in air pollution modelling are gamma, lognormal, and Weibull. This section will go over five more distributions that were used in this study. Table 1 presents the probability distributions used in this study. For estimating the parameters, the method of moments will be used.

2.3 Performance indicators

To compare the results for each set of predicted values using different statistical distributions, five performance measures were employed. Root mean square error (RMSE), normalized absolute error (NAE), coefficient of determination (R^2), index of agreement (IA), and prediction accuracy (PA) are the performance measures that have been employed. The RMSE presented the model's error in actual size (Prasad, Gorai, & Goyal, 2014), while NAE is sensitive in measuring the forecast model's residual error (Shcherbakov *et*

Table 1. Distributions and parameter estimator

Distribution	Probability density function (pdf)	Parameter estimator
Weibull	$f(x)=\left(\frac{\mu}{\lambda}\right)\left(\frac{x}{\lambda}\right)^{\mu-1}\exp\left[-\left(\frac{x}{\lambda}\right)^{\lambda}\right]$	$\lambda=\left(\frac{s^2}{\bar{x}}\right)^{-1.0852}$
	$x>0,\ \mu>0,\ \lambda>0$	$\mu=\frac{\bar{x}}{\Gamma\left(1+\frac{1}{\lambda}\right)}$
	(Forbes, Evans, Hasting, & Peacock, 2011)	(Kottegoda & Rosso, 1998)
Gamma	$f(x)=\left[\frac{1}{\mu\Gamma(\lambda)}\right]\left(\frac{x}{\mu}\right)^{\lambda-1}\exp\left(-\frac{x}{\mu}\right)$	$\mu=\frac{s^2}{\bar{x}}$
	$x>0,\ \mu>0,\ \lambda>0$	$\lambda=\left[\frac{\bar{x}}{s}\right]^2$
	(Forbes, Evans, Hasting, & Peacock, 2011)	(Forbes, Evans, Hasting, & Peacock, 2011)
Lognormal	$f(x)=\frac{1}{\sqrt{(2\pi\sigma^2)\lambda}}\exp\left[-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right]$	$\lambda=\sqrt{\ln\left[s^2+(\bar{x})^2\right]-2\ln(x)}$
		$\mu=\ln(\bar{x})-\frac{\lambda^2}{2}$
	$x>0,\ \mu>0,\ \lambda>0$	(Abdul Hamid, Yahaya, Ramli, & Ul-Saufie, 2013)
	(Zwillinger & Kokoska, 2000)	
Nakagami	$f(x)=\frac{2\lambda^{\lambda}}{\Gamma(\lambda)\mu^{\lambda}}x^{2\lambda-1}\exp\left(-\frac{\lambda}{\mu}x^2\right)$	$\lambda=\frac{E(x^2)}{E(x^2-\mu)^2}$
	$\mu=\frac{\delta}{s^2}$	$\mu=E(x^2)$
	(Binoti, Binoti, Leiti, Fardin, & Oliveira, 2012)	(Noga & Studanski, 2016)
Inverse Gaussian	$f(x)=\left[\frac{\mu}{2\pi x^3}\right]^{0.5}\exp\left[-\frac{\mu(x-\delta)^2}{2\delta^2 x}\right]$	$\mu=\frac{\delta}{s^2}$
	(Forbes, Evans, Hasting, & Peacock, 2011)	$\delta=\bar{x}$
		(Forbes, Evans, Hasting, & Peacock, 2011)
Log-logistic	$f(x)=\frac{\exp\left(\frac{\ln x-\mu}{\sigma}\right)}{\sigma\left[1+\exp\left(\frac{\ln x-\mu}{\sigma}\right)\right]^2}$	$\sigma=\bar{x}$
		$\mu=s^2$
	$x\geq 0,\ \mu>0,\ \sigma>0$	(Sharma, Sharma, Jain, & Kumar, 2013)
	(Sharma, Sharma, Jain, & Kumar, 2013)	
Logistic	$f(x)=\frac{\operatorname{sech}^2\left[\frac{x-\delta}{2\mu}\right]}{4\mu}$	$\delta=\bar{x}$
		$\mu=\sqrt{\frac{3s}{\pi^2}}$
	(Forbes, Evans, Hasting, & Peacock, 2011)	(Forbes, Evans, Hasting, & Peacock, 2011)
Birnbaum-Saunders	$f(x)=\frac{1}{2\lambda\mu\sqrt{2\pi}}\left[\left(\frac{\mu}{x}\right)^{0.5}+\left(\frac{\mu}{x}\right)^{1.5}\right]\exp\left[-\frac{1}{2\lambda^2}\left(\frac{x}{\mu}+\frac{\mu}{x}-2\right)\right]$	$\lambda=\left\{2\left[\left(\frac{\bar{x}}{r}\right)^{0.5}-1\right]\right\}^{0.5}$
	(Thonglim, Budsaba, & Volodin, 2014)	$\mu=0.5(\bar{x}.r)$
		$r=\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right)$
		(Thonglim, Budsaba, & Volodin, 2014)

al., 2013). The R^2 reflects the degree of model error fitting (Rybarczyk & Zalakeviciute, 2018), the AI determines the magnitudes of the actual values in relation to the model's predicted value (Prasad *et al.*, 2014) and the PA was utilized as an indicator to determine the estimator performance of the generated model (Junninen, Niska, Tuppurainen, Ruuskanen, & Kolehmainen, 2004). The formulae for each performance indicator that has been used are given in Table 2.

where *n* is the number of monitoring records, *Oi* is the observed monitoring records, and *Pi* is the predicted values. Smaller values of RMSE and NAE are better, while for R_2, IA and PA, a value closer to 1 indicates a better estimator.

3. Results and Discussion

3.1 Descriptive statistics

Table 3 shows the descriptive statistics for hourly average data from Kuala Terengganu, Jerantut, and Banting

Table 2. Performance indicators

$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2}$
$NAE = \frac{\sum_{i=1}^{n}\lvert P_i - O_i\rvert}{\sum_{i=1}^{n} O_i}$
$R^2 = \left[\frac{1}{n}\frac{\sum_{i=1}^{n}(P_i - \overline{P})(O_i - \overline{O})}{\sigma_P \sigma_O}\right]^2$
$IA = 1 - \left[\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(\lvert P_i - \overline{O}\rvert + \lvert O_i - \overline{O}\rvert)^2}\right]$
$PA = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(P_i - \overline{P})(O_i - \overline{O})}{\sigma_P \sigma_o}$

Table 3. Summary of descriptive statistics

Year	2017	2018	2019	2020
Kuala Terengganu				
Mean	0.0176	0.0174	0.0159	0.0141
Std Deviation	0.0119	0.0125	0.0112	0.0098
Minimum	0.0001	0.0001	0.0001	0.0001
Maximum	0.0218	0.0694	0.0579	0.0570
Skewness	0.5130	0.6350	0.5200	0.5210
Jerantut				
Mean	0.0152	0.0163	0.0189	0.0120
Std Deviation	0.0104	0.0117	0.0131	0.0105
Minimum	0.0001	0.0001	0.0002	0.0003
Maximum	0.0546	0.0595	0.0668	0.0609
Skewness	0.7290	0.7020	0.6080	0.6210
Banting				
Mean	0.0187	0.0198	0.0195	0.0177
Std Deviation	0.0150	0.0195	0.0190	0.0171
Minimum	0.0001	0.0001	0.0001	0.0003
Maximum	0.1280	0.1028	0.1011	0.0928
Skewness	1.1230	0.9940	0.8780	1.0500

monitoring stations from 2017 to 2020. The highest concentrations were observed in 2017 at the Banting monitoring station, with a reading of 0.1280 ppm. Jerantut is a reference monitoring station located far from urban centres. However, the maximum reading and mean value for this station do not differ considerably from those recorded in Kuala Terengganu. According to Department of Environment Malaysia (DoE, 2014) the highest O_3 concentration occurred due to high traffic volume and a conducive atmospheric condition for O_3 formation. Nevertheless, during this study period, the ground-level ozone in Kuala Terengganu and Jerantut monitoring station did not exceed the limit set by MAAQG, which was 0.10 ppm. However, the Banting monitoring station shows that the maximum reading of O_3 concentration each year exceeded the 0.1000 ppm limit. This may be due to the location being nearer to the Kuala Lumpur International Airport (KLIA) and the Genting Sanyen power plant, which may affect the O_3 concentration.

The bar charts in Figure 2 show the means of ground-level ozone concentration in Kuala Terengganu, Jerantut, and Banting monitoring stations in 24 hours, from 2017 to 2020. Kuala Terengganu showed the occurrences of the highest concentrations at 3.00 pm, while at 7.00 am the lowest concentrations were recorded. The ground-level ozone started to increase at 8:00 am until it reached a peak at 3.00 pm, and later slowly began to decrease simultaneously as the sunlight intensity decreased. In Jerantut, the occurrences of the highest concentrations also appeared at 3.00 pm, while at 7.00 am the lowest concentrations were observed. Banting monitoring station showed the highest average reading of O_3 concentration, surpassing Kuala Terengganu and Jerantut. It reached the highest concentrations at 3.00 pm and the lowest concentrations recorded were at 1.00 am. The highest concentration of ozone was caused by the intense sunlight and the presence of nitrogen dioxide that reacted in the sunlight (Geng *et al.*, 2018).

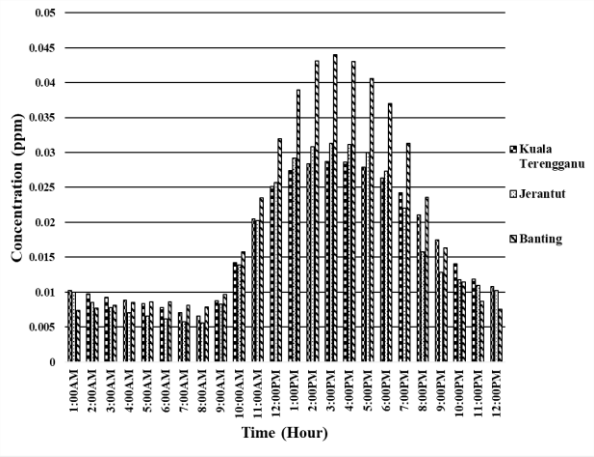

Figure 2. Behavior of the O_3 concentration at all monitoring stations

3.2 Distribution fitting

For Kuala Terengganu monitoring station, Table 4 clearly shows that the Nakagami distribution has a better fit than the other distributions, for all the years involved in this study. Five performance indicators were used for comparing those distributions to obtain the best distribution. The performance indicators involved are root mean square error (RMSE), normalized absolute error (NAE), coefficient of determination (R^2), index of agreement (IA) and prediction accuracy (PA). The performance indicators consistently indicate that the Nakagami distribution yields low values for error measures and high values for accuracy measures. However, it's worth noting that in the year 2018, the gamma distribution outperformed the Nakagami distribution in one specific indicator, which is the R^2.

The probability density function (pdf) and cumulative distribution function (cdf) plots in Figure 3 and Figure 4 show the distribution curve of O_3 concentration for Kuala Terengganu monitoring station. The pdf plotted is a positive skew distribution. This indicates that most of the O_3 concentration readings are low and centered on the left side of the distribution. Moreover, the cdf was plotted based on the best distribution that fit the O_3 concentration. The cdf plot

Table 4. The performance indicators for comparing distributions for Kuala Terengganu

Year	PI	Log-Normal	Gamma	Weibull	Nakagami	Birnbaum-Saunders	Inv. Gaussian	Logistic	Loglogistic
2017	NAE	0.3561	0.1247	58.9600	0.0752	0.3134	0.4206	0.1534	1.0051
	RMSE	0.0164	0.0032	1.3734	0.0027	0.0090	0.0147	0.0046	4.6854
	IA	0.8084	0.9833	0.0273	0.9888	0.9156	0.8288	0.9675	0.2757
	PA	0.8521	0.9674	0.9518	0.9779	0.9055	0.8426	0.9396	0.4631
	R^2	0.7260	0.9356	0.9057	0.9562	0.8198	0.7098	0.8828	0.2145
2018	NAE	0.2800	0.1031	57.1137	0.0491	0.2368	0.4030	0.1249	0.4630
	RMSE	0.0140	0.0021	1.3733	0.0010	0.0061	0.0145	0.0036	0.0327
	IA	0.8695	0.9922	0.0264	0.9982	0.9535	0.8269	0.9791	0.5693
	PA	0.8960	0.9847	0.9628	0.9966	0.9525	0.8503	0.9622	0.7144
	R^2	0.8027	0.9695	0.9267	0.9930	0.9070	0.7229	0.9256	0.5103
2019	NAE	0.0299	0.1171	63.5063	0.0686	0.1988	0.4493	0.1163	0.5375
	RMSE	0.6369	0.0021	1.3755	0.0013	0.0050	0.0145	0.0031	0.0338
	IA	0.8034	0.9899	0.0232	0.9964	0.9717	0.7955	0.9805	0.5140
	PA	0.6453	0.9802	0.9546	0.9933	0.9698	0.8252	0.9642	0.6855
	R^2	0.0299	0.9606	0.9111	0.9865	0.9404	0.6809	0.9295	0.4698
2020	NAE	0.2965	0.1209	69.5643	0.0713	0.1844	0.4518	0.1171	0.4818
	RMSE	0.0130	0.0020	1.3773	0.0012	0.0052	0.0130	0.0027	0.0264
	IA	0.8710	0.9891	0.0206	0.9957	0.9743	0.7931	0.9809	0.5612
	PA	0.9063	0.9786	0.9503	0.9919	0.9741	0.8248	0.9651	0.7102
	R^2	0.8212	0.9576	0.9029	0.9837	0.9487	0.6802	0.9312	0.5043

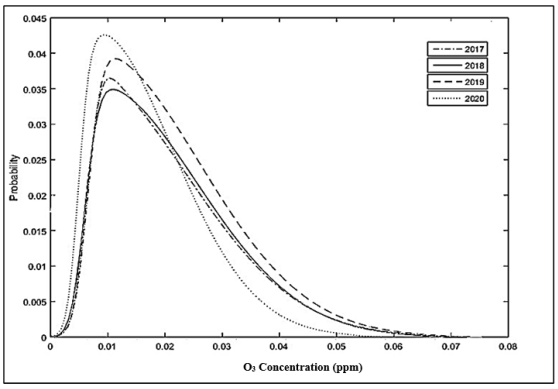

Figure 3. Probability density function plot of ground-level ozone for Kuala Terengganu

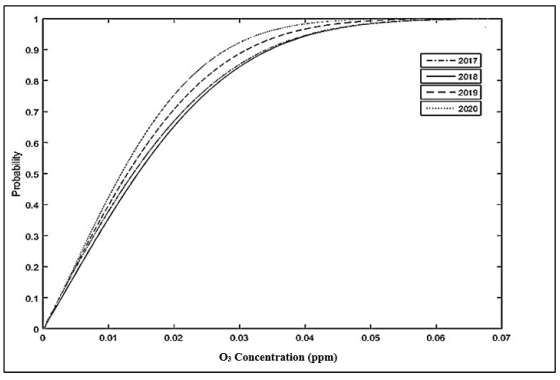

Figure 4. Cumulative distribution function plot of ground-level ozone for Kuala Terengganu

shows that the probability of the O_3 concentration to exceed the limit is zero. Since the probability that O_3 concentration does not exceed the MAAQG limit, the air in this area is still not contaminated by O_3 pollutant.

The results show that the probability of concentration exceeding the limit 0.10 ppm was 0 [that is, $P(X>0.10) = 0$] for all the years involved in this study. This indicates that the ground-level ozone in Kuala Terengganu monitoring station for the entire year stayed at a concentration below 0.10 ppm. Therefore, there is no return period predicted for concentrations above the limit set by MAAQG.

The results for Jerantut and Banting also show that the Nakagami distribution fit the data better compared than the other types of distributions. Table 5 and Table 6 show the performance of each distribution for the data from Jerantut and Banting monitoring stations, respectively.

Figure 5 to Figure 8 show the pdf and cdf of the best distribution for Jerantut and Banting. For Jerantut, the pdf shows a positively skewed curve and the pdf plot for 2018 and 2019 show longer tails than in the other years. This is sign of higher O_3 concentration peaks compared to 2017 and 2020. Even though the O_3 concentrations do not exceed the 0.1 ppm limit, caution is needed in this area to control the future increases in concentration. In the cdf plot, no occurrences exceed the 0.1 ppm limit in this area, indicating that Jerantut area still had a good air quality level not contaminated by O_3. Hence, there is no return period that could be predicted for Jerantut station.

The O_3 concentrations from 2017 to 2020 in Banting in the pdf plot have a longer tail, which indicates that there are more extreme events. Moreover, the trends show that the O_3 concentration exceeded the limit 0.10 ppm. Furthermore, the cdf plot shows that the O_3 concentration tends to exceed the 0.10 ppm limit. For instance in 2017 and 2018, the predicted values of the concentration exceeded 0.11ppm. The yearly increases of O_3 concentration depend on many factors, led by an increase in vehicles that produce nitrogen oxides (NOx), CO, and VOCs. This indicates that this area was contaminated with the O_3 pollutant every year.

The predictions of exceedances for the ground-level ozone are based on the best fit distributions. Moreover, the value for the exceedances is taken from the cdf plots of the

Table 5. The performance indicators for comparing distributions for Jerantut

Year	PI	Log-Normal	Gamma	Weibull	Nakagami	Birnbaum-Saunders	Inv. Gaussian	Logistic	Loglogistic
2017	NAE	0.1848	0.0491	67.4556	0.0413	0.2360	0.2955	0.1298	1.0037
	RMSE	0.0079	0.0011	1.3768	0.0007	0.0060	0.0089	0.0029	4.6178
	IA	0.9010	0.9966	0.0211	0.9983	0.9366	0.8828	0.9767	0.2186
	PA	0.9224	0.9935	0.9757	0.9970	0.9493	0.9096	0.9599	0.6184
	R^2	0.8506	0.9869	0.9518	0.9938	0.9011	0.8272	0.9213	0.3823
2018	NAE	0.2392	0.0867	61.6009	0.0319	0.1391	0.3841	0.1393	0.3901
	RMSE	0.0115	0.0019	1.3746	0.0008	0.0035	0.0133	0.0036	0.0271
	IA	0.8769	0.9928	0.0250	0.9987	0.9820	0.8318	0.9762	0.6090
	PA	0.9066	0.9860	0.9666	0.9974	0.9850	0.8539	0.9579	0.7235
	R^2	0.8217	0.9720	0.9341	0.9946	0.9701	0.7290	0.9173	0.5233
2019	NAE	0.2242	0.0922	52.7009	0.0382	0.2449	0.3888	0.1207	0.3899
	RMSE	0.0109	0.0022	1.3719	0.0010	0.0080	0.0152	0.0037	0.0301
	IA	0.8931	0.9922	0.0277	0.9983	0.9284	0.8267	0.9795	0.6133
	PA	0.9225	0.9847	0.9606	0.9968	0.9371	0.8506	0.9634	0.7283
	R^2	0.8509	0.9695	0.9225	0.9935	0.8780	0.7234	0.9279	0.5303
2020	NAE	0.1660	0.0927	65.9176	0.0404	0.2325	0.3299	0.1244	0.3652
	RMSE	0.0084	0.0017	1.3762	0.0007	0.0079	0.0101	0.0030	0.0219
	IA	0.9259	0.9926	0.0222	0.9985	0.9339	0.8661	0.9792	0.6482
	PA	0.9381	0.9855	0.9619	0.9971	0.9457	0.8765	0.9627	0.7491
	R^2	0.8798	0.9711	0.9251	0.9941	0.8941	0.7681	0.9266	0.5611

Table 6. The performance indicators for comparing distributions for Banting

Year	PI	Log-Normal	Gamma	Weibull	Nakagami	Birnbaum-Saunders	Inv. Gaussian	Logistic	Loglogistic
2017	NAE	0.6090	0.1151	49.1085	0.0485	0.3465	0.6555	0.2830	1.0086
	RMSE	0.0478	0.0029	1.3659	0.0012	0.0117	0.0349	0.0082	4.9554
	IA	0.6235	0.9941	0.0440	0.9990	0.9371	0.7078	0.9547	0.3269
	PA	0.7777	0.9887	0.9889	0.9985	0.9317	0.7661	0.9222	0.2949
	R^2	0.6048	0.9774	0.9778	0.9967	0.8679	0.5869	0.8502	0.0869
2018	NAE	0.4188	0.1437	50.8382	0.0710	0.1702	0.7700	0.2704	2.0052
	RMSE	0.0285	0.0035	1.3670	0.0018	0.0040	0.0414	0.0077	0.2750
	IA	0.6746	0.9910	0.0419	0.9977	0.9409	0.6263	0.9584	0.1163
	PA	0.7247	0.9826	0.9825	0.9956	0.9123	0.6997	0.9270	0.4618
	R^2	0.5251	0.9653	0.9652	0.9911	0.8321	0.4895	0.8591	0.2132
2019	NAE	0.3293	0.1728	51.5120	0.0981	0.3940	0.7679	0.2797	2.0132
	RMSE	0.0167	0.0043	1.3675	0.0025	0.0117	0.0403	0.0077	0.2750
	IA	0.8037	0.9862	0.0410	0.9952	0.7869	0.6185	0.9572	0.1079
	PA	0.8397	0.9732	0.9729	0.9907	0.7861	0.6761	0.9246	0.4322
	R^2	0.7050	0.9469	0.9464	0.9813	0.6178	0.4571	0.8547	0.1868
2020	NAE	0.1652	0.1257	56.3739	0.0581	0.4273	0.5007	0.2631	1.0368
	RMSE	0.0066	0.0028	1.3699	0.0011	0.0218	0.0215	0.0068	0.1083
	IA	0.9375	0.9926	0.0370	0.9988	0.7569	0.7989	0.9579	0.2727
	PA	0.9319	0.9857	0.9845	0.9980	0.7537	0.8183	0.9266	0.5627
	R^2	0.8682	0.9714	0.9691	0.9958	0.5680	0.6695	0.8585	0.3165

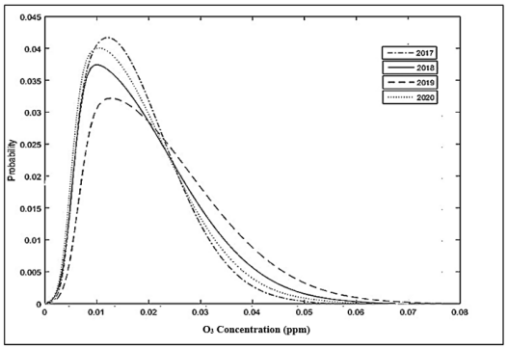

Figure 5. Probability density function plot of ground-level ozone for Jerantut

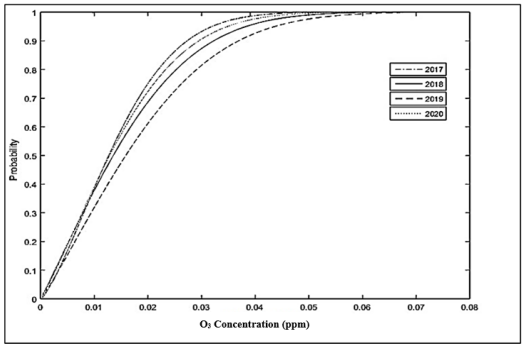

Figure 6. Cumulative distribution function plot of ground-level ozone for Jerantut

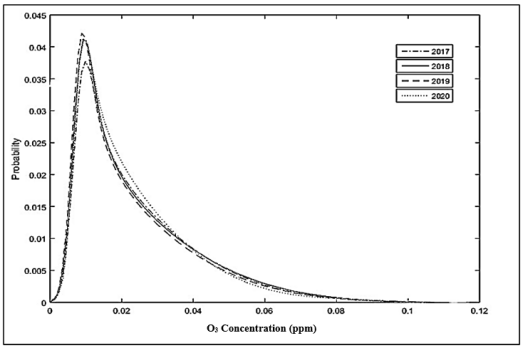

Figure 7. Probability density function plot of ground-level ozone for Banting

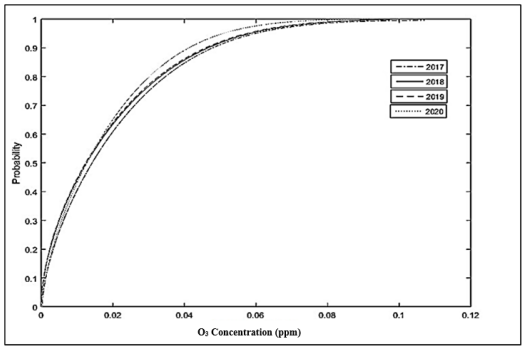

Figure 8. Cumulative distribution function plot of ground-level ozone for Banting

best distribution. As shown in the table below, the best fit distribution is Nakagami. Table 7 shows the actual return period and predicted return period. The results indicate the probability that concentration exceeded the limit of 0.10 ppm [that is, P ($X > 0.10$) $\neq 0$] from the year 2017 until 2019. This indicates that the ground-level ozone in Banting monitoring station has reached a concentration exceeding 0.10 ppm. Furthermore, it shows there is evidence of return period, predicted for concentration to exceed the limit set by MAAQG.

Table 7. The actual return period and predicted return period

Year	Actual exceedance (Days)	Predicted exceedance (Days)	Actual return period (Days)	Predicted return period (Days)
2017	12	13	730	676
2018	5	12	1738	730
2019	1	9	8900	973
2020	0	0	0	0

4. Conclusions

Environmental pollution, especially ozone pollution, should be taken into consideration especially if it occurs daily, as that may give negative effects on human health and the ecosystem itself. The O_3 concentration for Kuala Terengganu has a lower average compared to the Banting monitoring station, remaining below 0.10 ppm. According to the findings of this study, when assessed based on air pollution monitoring areas, Banting, a sub-urban region, has exhibited notably higher air pollution readings. On multiple occasions, these readings have exceeded the thresholds established by the Malaysia Ambient Air Quality Guidelines. The expected O_3 concentrations for the Banting station are influenced by the surroundings and the amount of NO_X. Banting monitoring station is near the Genting Sanyen power plant and the Kuala Lumpur International Airport (KLIA), which contribute as sources of the NO_X. Meteorological factors such as temperature, UV radiation, and wind speed also affect the concentration levels. Based on the performance indicators, the Nakagami distribution, which was first employed in air pollution modelling, was the best fit for all three monitoring sites from 2017 to 2020. As only the Banting monitoring station had readings of ground level ozone that exceed the guideline, the return period for this station has also been predicted by using the Nakagami distribution to estimate when there will be a recurrence of ground level ozone readings exceed the limit.

Acknowledgements

The author would like to thank Ministry of Education and Universiti Sains Malaysia for the funding (FRGS/1/2019/STG06/USM/02/7) and Universiti of Tun Hussein Onn, Malaysia (UTHM) for GPPS grant (Vote U700).

References

Abdul Hamid, H., Yahaya, A. S., Ramli, N. A., & Ul-Saufie, A. Z. (2013). Finding the best statistical distribution model in PM_{10} concentration modeling by using lognormal distribution. *Journal of Applied Sciences*, *13*(2), 294-300. doi:10.3923/jas.2013.294.300

Abdullah, A. M., Ismail, M., Yuen, F. S., Abdullah, S., & Elhadi, R. E. (2017). The relationship between daily maximum temperature and daily maximum ground level ozone concentration. *Polish Journal Environmental Studies*, *26*(2), 517–523. doi:10. 15244/pjoes/65366

Ahmad Isiyaka, H., Juahir, H., Toriman, M. E., Gasim, B. M., Azid, A., Amri, M. K., . . . Garba, M. A. (2014). Spatial assessment of air pollution index using environmental modeling technique. *Advances in Environmental Biology*, *8*(24), 244-256.

Awang, N. R., & Ramli, N. A. (2017). Preliminary study of ground level ozone nighttime removal process in an urban area. *Journal of Tropical Resources and Sustainable Science*, *5*(2), 83–88. doi:10.47253/ jtrss.v5i2.595

Binoti, D. H. B., Binoti, M. L. M. S., Leite, H. G., Fardin, L., & Oliveira, J. C. (2012). Probability density functions for description of diameter distribution in thinned stands of Tectona grandis. *Cerne*, *18*(2), 185-196. doi:10.1590/S0104- 77602012000200002

Bolsoni, V. P., Oliveira, D. P., Pedrosa, G. S., & Souza, S. R. (2018). Volatile organic compounds (VOC) variation in Croton floribundus (L.) Spreng. related to environmental conditions and ozone concentration in an urban forest of the city of São

Paulo, São Paulo State, Brazil. *Hoehnea*, *45*(2), 184-191. doi:10.1590/2236-8906-60/2017.

Department of Environment, Malaysia. (2014). Malaysia environmental quality report 2014. Kuala Lumpur, Malaysia: Ministry of Science, Technology and the Environment.

Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2011). *Statistical distributions*. Toronto, Canada: John Wiley and Sons.

Geng, F., Tie, X., Xu, J., Zhou, G., Peng, L., Gao, W., . . . Zhao, C. (2018). Characterizations of ozone, NOx, and VOCs measured in Shanghai, China. *Atmospheric Environment*, *42*(29), 6873–6883. doi:10.1016/j.atmosenv.2008.05.045

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Method for imputation of missing values in air quality data sets. *Atmospheric Environment*, *38*(18), 2895-2907. doi:10.1016/j.atmosenv.2004.02.026

Kim, K. H., Kabir, E., & Kabir, S. A (2015). A review on the human health impact of airborne particulate matter. *Environment International*, *74*,136–143. doi:10. 1016/j.envint.2014.10.005.

Kottegoda, N. T., & Rosso, R. (1998). *Statistics, probability and reliability for civil and environmental engineer*. Singapore: McGraw-Hill.

Latif, M. T., Huey, L. S., & Juneng, L. (2012). Variations of surface ozone concentration across the Klang Valley, Malaysia. *Atmospheric Environment*, *61*, 434-445. doi:10.1016/j.atmosenv.2012.07.062.

Mabahwi, N. A., Ling, H. L., & Omar, D. (2014). Human health and wellbeing: Human health effect of air pollution. *Procedia - Social and Behavioral Sciences 153*, 221-229. doi:10.1016/j.sbspro.2014. 10.056.

Maciejewska, K., Rezlar, K. J., Reizer, M., & Klejnowski, K. (2015). Modelling of black carbon statistical distribution and return periods of extreme concentrations. *Environmental Modelling and Software*, *74*, 212-226. doi:10.1016/j.envsoft.2015. 04.016

Mohamad, N. S., Deni, S. M. & UI-Saufie, A. Z. (2018). Application of the first order of Markov Chain model in describing the PM_{10} occurrences in Shah Alam and Jerantut, Malaysia. *Pertanika Journal of Science and Technology*, *26*(1), 367- 378.

Noga, K. M., & Studanski, R. (2016). Estimation of Nakagami distribution parameters in describing a fading radio-communication channel. *Maritime Technical Journal*, *204*(1), 69–81. doi:10.5604/0860889x.1202 437

Prasad, L., Gorai, A. K., & Goyal, P. (2016). Development of ANSIS models for air quality forecasting and input optimization for reducing the computational cost and time. *Atmospheric Environment*, *128*, 246-262. doi:10.1016/j.atmosenv.2016.01.007

Rybarczyk, Y., & Zalakeviciute, R. (2018). Machine learning approaches for outdoor air quality modelling: A systematic review. *Applied Sciences*, *8*(12), 1-28. doi:10.3390/app8122570

Sharma, P., Sharma, P., Jain S., & Kumar, P. (2013). An integrated statistical approach for evaluating the exceedances of criteria pollutant in the ambient air of megacity Delhi. *Atmospheric Environment*, *70*, 413-414. doi:10.1016/j.atmosenv.2013.02.021

Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., & Kamaev, V. A. (2013). A survey of forecast error measure. *World Applied Sciences Journal*, *24*(24), 171-176. doi:10.5829/idosi.wasj.2013.24.itmies.80032

Song, X., & Hao, Y. (2022). Analysis of ozone pollution characteristics and transport paths in Xi'an city. *Sustainability*, *14*(23), 16146. doi:10.3390/su1423 16146.

Suparta, Y. W., Alhasa, K. M., Singh, M. S. J., & Latif, M. T. (2015). The development of PWV index for air pollution concentration detection in Banting, Malaysia. *Proceedings of the International Conference on Space Science and Communication*, 498-502. doi:10.1109/IconSpace.2015.7283810.

Thonglim, P., Budsaba, K., & Volodin, A. I. (2014). Asymptotic confidence ellipses of parameters for the Birnbaum-Saunders distribution. *Thailand Statistician*, *12*(2), 207-222.

Zaki, T. N. A. M., Faizah, N., Yusof, F., & Shith, S. (2016). Morphology analysis of fine particles in background station of Malaysia. *Sustainability in Environment*, *1*(1), 12-24. doi:10.22158/se.v1n1p12

Zeng, P., Lyu, X. P., Guo, H., Cheng, H. R., Jiang, F., Pan, W. Z., . . . Hu, Y. Q. (2018). Causes of ozone pollution in summer in Wuhan, Central China. *Environmental Pollution*., *241*, 852–861. doi:10.1016/j.envpol. 2018.05.042

Zhang, Y., Qu, S., Zhao, J., Zhu, G., Zhang, Y., Lu, X., . . . Wang, H. (2018). Quantifying regional consumption-based health impacts attributable to ambient air pollution in China. *Environment International*, *112*, 100-106. doi:10.1016/j.envint. 2017.12.021

Zhou, Q., Cheng, L., Zhang, Y., Wang, Z., & Yang, S. (2022). Relationships between springtime $PM_{2.5}$, PM_{10} and O_3 pollution and the boundary layer structure in Beijing, China. *Sustainability*, *14*(15), 9041. doi:10. 3390/su14159041.

Zwillinger, D., & Kokoska, S. (2000). *CRC standard probability and statistics tables and formulae*. New York, NY: Chapman and Hall.